Pose Transformers: Human Motion Prediction with Non-Autoregressive Transformers

Angel Martínez-González, Michael Villamizar and Jean-Marc Odobez

Motivations & Contributions

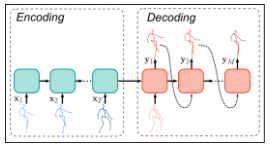
Motivations

- Activity recognition & behavior understanding

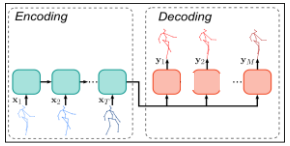
SOTA

- Autoregressive decoding (AR)
  - computational costly; error propagation

AR decoding



Parallel decoding



Contributions

- Non-autoregressive Transformer architecture
  - parallel inference can lead to efficiency
- Joint activity and motion prediction
  - Activity may allow better model selection

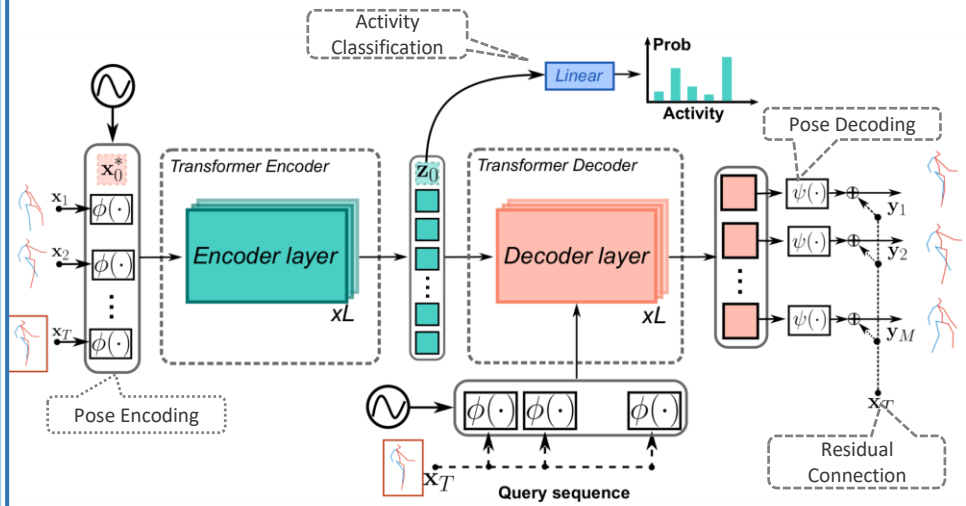
Related Work

[1] Martinez et al, On Human motion prediction with recurrent neural networks, CVPR 17
[2] Carion et al, Detection Transformers, ECCV 2020
[3] Dosovitskiy et al, Vision Transformers, ICLR 2021
[4] Mao et al, History repeats itself, ECCV 2020
[5] Gu et al, Non-Autoregreesive Machine Translation, ICRL 2018

Acknowledgements

MuMMER mummer-project.eu
CONACYT Consejo Nacional de Ciencia y Tecnología

Method

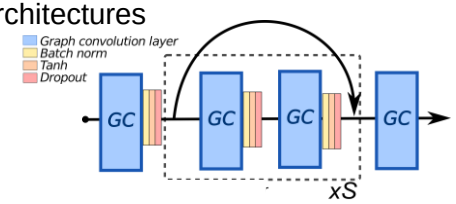


- Query sequence: repeat x_T
  - Parallel motion decoding
- Residual pose decoding
- Activity encoding – learnable activity token

Pose Encoding & Decoding

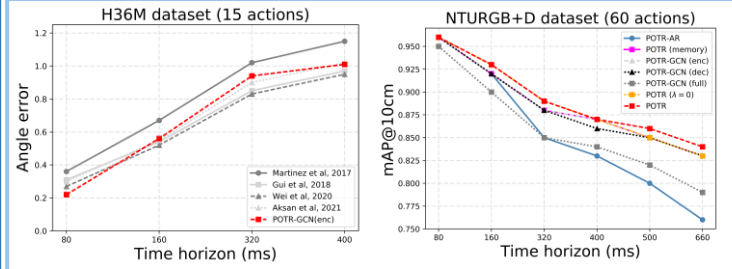
Investigated architectures: ϕ and ψ architectures

1. Linear layers
2. Graph Convolutional Network



Code & Models
https://github.com/idiap/potr

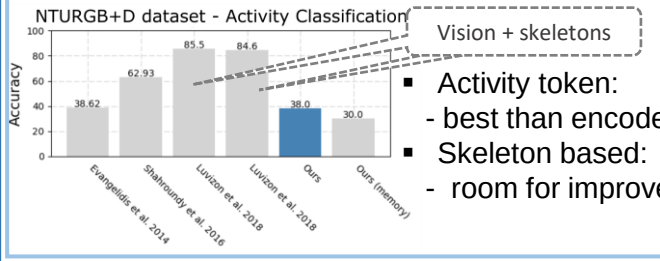
Motion Prediction



- Ours – best in short-term horizons
- Speed - **Non-AR: 149.2 SPS***; AR: 8.9 SPS

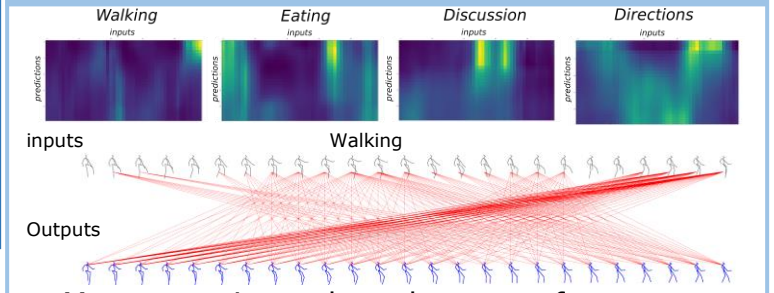
*SPS: Sequences per Second

Activity Classification



- Activity token:
  - best than encoder maps
- Skeleton based:
  - room for improvement

Attention Visualization



- More attention to last elements of sequence
- Attention in cyclic patterns of motion